“模型落地”和“语境转换”
——数字技术赋能下的历史研究

陈　松
(巴克内尔大学 东亚系, 美国)

摘　要: 历史研究既离不开史料,也离不开作为历史学家思维建构的模型。模型对过去发生的事件加以简化和抽象,给过去赋予意义。如何跨越横亘于史料与模型之间的鸿沟,是所有历史学家都必须面对的问题。数字技术凭借其强大的数据处理能力,为历史学家将模型落实于史料之中提供了前所未有的可能。数字技术的应用也深刻改变了历史学家与史料互动的方式。它引导历史学家从单篇文献的固有语境中跳脱出来,基于史料之间内在的呼应关系、通过数据提取和数据关联搭建起更为宏观、更为结构化的新语境,进而揭示在文献细读时不易察觉的历史现象。数字技术拓宽了史学研究的视野,也拓展了史学研究的范式,鼓励历史学家将个案研究与整体性研究相结合,从更加全局性的、统合的视角重新审视个别现象的历史意义。

关键词: 模型落地;语境转换;量化历史;数字史学;数字人文

中图分类号: G 250. 7　　**文献标识码:** A　　**文章编号:** 1000-5919(2024)06-0100-13

近年来,数字技术的迅猛发展正在深刻地改变人文学科的研究面貌。来自历史、文学、社会科学、计算机科学等诸多领域的学术团队不仅借助数字化手段保存和分享历史资料,也大力发展数字基础设施,如建设数据库、开发便于人文学者使用的分析工具、搭建融合学术研究与知识普及为一体的智慧数据平台与可视化平台、重建历史场景、探索数字记忆和数字叙事。① 借助数字技术,越来越多的学者通过数据挖掘、文本分析等方法,揭示出传统研究难以发现的规律和特征,大幅提升了研究效率并拓宽了学术视野。数字技术与人文研究的融合类型多样、层次丰富,也催生了对数字人文的“大帐篷”定义。这一极具包容性的定义旨在涵盖各种处于数字技术和人文学科交叉地带的研究方法和课题,既包括利用数字工具开展人文研究,也包括使用人文学科的视角分析数字社会的现象,对数字化社会中的人文问题进行反思,既包括通过数字技术创造新知识,也包括利用数字手段推广和普及现有知识的努力。“大帐篷”定义凸显了数字人文领域的多样性与灵活性,包容了不同形式的研究与探索。以 2024 年夏季在华盛顿举办的数字人文大会为例,会议论文中最常用的关键词涵盖了从机器学习、自然语言处理、人工智能、计算机视觉,到数字史学、计算文学、文本分析、社会网络分析,再到数字典藏、数字出版、命名实体识别、开放关联数据、基础设施和可视化等诸多领域。数字人文的学术内涵与研究路径如此多姿多彩,自然也从不同层面、不同角度推动了人文学科的研究和教学。② 要面面俱到地讨论数字技术与人文研究的关系自无可能。因此,本文主要从笔者本人的研究经验出发,从“模型落地”和“语境转换”两个角度,探讨数字技术如何改变了历史学家与史料的互动方式,帮助历史学家从全新的视角提出并解答研究

收稿日期: 2024-04-09

作者简介: 陈松,巴克内尔大学东亚系中国史副教授。

① 浙江大学智慧古籍平台(https://csab.zju.edu.cn)。学术地图发布平台(http://amap.zju.edu.cn)。唐宋文学编年地图(https://cnkgraph.com/Map/PoetLife)。冯惠玲:《数字人文视角下的数字记忆——兼议数字记忆的方法特点》,《数字人文研究》2021 年第 1 期,第 87—95 页。

② “DH2024 Submissions,” 由 apjanco 于 2023 年 12 月 11 日发布。https://dh2024.adho.org/dh2024—submissions/.

问题。

模型落地

从最抽象的层面说,数字技术凭借其强大的数据挖掘、数据分析能力,为历史学家跨越模型与实证证据之间的鸿沟提供了强大的技术支持,推动了“模型落地”。

什么是模型? 要回答这个问题,首先需要反思历史的本质。卡尔·贝克尔(Carl L. Becker)曾指出,历史有两层含义:一是曾经发生过的真实事件(actual series of events),二是经我们认定而成为我们记忆之一部分的一连串观念(ideal series)。我们不妨将它们分别称作历史实态和历史知识。过去发生的事件是绝对的、不可改变的,但它们已随时间而逝,今人所知的历史只是我们依赖史料对过去所作的推断。历史事实潜藏于史料之中,只有当历史学家将相关记述挑选出来并加以认定时,它们才成为我们观念中的“事实”,成为我们的历史知识 ①。不过,通常情况下,历史学家的工作不会止步于挑选和确认。正如大卫·罗温素(David Lowenthal)所言,要想理解过去,就必须重新讲述它,而这个过程离不开讲述者的创造性。“重新讲述”意味着对历史记述进行“归并、夸大、精简”,将举足轻重、独一无二的部分凸显出来,而将一成不变、细枝末节的部分进行淡化处理。这么做的目的并非为了改变或扭曲历史人物和事件,而是为了赋予它们历史意义,使它们栩栩如生。简言之,我们所知的历史不是过去发生了什么,而是后人对它的理解,是后人借由自己的想象力对过去做出的解读。②

因此,历史知识既不能脱离史料而存在,也不能脱离历史学家的思维建构(intellectual construct)而存在。舍去思维建构,就无法给过去赋予历史意义,解读亦无从谈起。彼得·伯克(Peter Burke)将这种思维建构称为“模型”(model),它“对现实加以简化,以便理解现实……旨在凸显那些反复出现的、具有普遍性和典型性的事物”。在伯克看来,尽管排斥模型的历史学家不乏其人,但是事实上他们中“大多数人都在使用模型,只不过……没有意识到自己是在这么做”。历史学家口中的“革命”“封建制”“文艺复兴”“农民经济”,凡此种种,无一不是模型。脱离模型,历史学家就无法进行思考。③

对于历史学家来说,放在他们面前的是纷繁复杂的史料,他们要提出的研究成果则是对过去进行简化、抽象的思维模型。历史学家总是从史料和理论假设出发,建立一种能够概括研究对象之特征的历史认知。这一过程,换言之,就是在定义不同的模型,并通过在史料中验证每个模型的适用性,对模型加以调整,逐步形成对历史的理解。历史知识的形成过程就是在模型与史料中互动的过程中,改进模型将其落实于历史证据的过程。可以说,所有历史研究都是在对过去建模(modeling of the past),所有历史学家都是在努力跨越横亘于史料与模型之间的鸿沟,所有历史知识都是在史料与模型的对接与拟合中创造出来的。

如果说史料是对历史实态的刻画,那么模型就是对历史实态的简化和抽象。模型与史料的拟合注定不会严丝合缝,但也正因如此,它才可能成为历史学家分析思考历史实态的工具。模型的定义有单征性(monothetic)和多征性(polythetic)之分。在单征性定义下,一个实例的归属资格取决于它是否完全具备用于定义该模型的一个或一组属性。多征性模型虽然也是由一系列属性来定义的,但是并非每个实例都完全具备这些属性,通常情况下,每个实例只会拥有其中绝大多数属性,而每个属性也为绝大多数

---

① 卡尔·贝克尔:《人人都是他自己的历史学家》,收入氏著《人人都是他自己的历史学家——论历史与政治》,北京:北京大学出版社 2013 年版,第 196、208—209 页。

② David Lowenthal, *The Past is a Foreign Country —— Revisited*, Cambridge, UK: Cambridge University Press, 2015, p. 341.

③ 彼得·伯克:《历史学与社会理论》,上海:上海人民出版社 2019 年版,第 44—45 页。本处引文对原译稍有调整。他将“模型”等同于马克斯·韦伯的理想型(ideal type),源于希腊语中的 typos。

实例所共享。① 伯克以保罗·维诺格拉多夫爵士与 F. W. 梅特兰关于中世纪欧洲庄园的论战为例,阐释了多征性模型的概念。维诺格拉多夫在其 1892 年的著作中为“普通庄园的结构”勾勒出一副经典图景,揭示出庄园普遍具有的一系列属性特征,如维兰、直领地、法庭等。他的观点遭到梅特兰的批判。梅特兰指出,“在维诺格拉多夫识别出的那组特征中,每一项特征都会在一些实例中付之阙如。有些庄园没有维兰,有些庄园没有自由佃农,有些没有直领地,有些没有法庭”,因此梅特兰断言“要描述一种典型的庄园是一项不可能完成的功绩”。对此,伯克敏锐地指出,从实证角度看,梅特兰的观点无疑是正确的,但梅特兰没有意识到维诺格拉多夫所描述的“庄园”其实只是一个多征性定义下的模型。② 这种模型是分析性概念,它不可能也不应该对历史实态进行巨细靡遗的描述。也正因如此,模型才能作为分析思考的工具,也才有其价值。

考虑到历史现象的复杂性以及史料的残缺不全,多征性定义为研究历史现象提供了灵活有效的分析框架。但随之而来的问题是,多征性模型的边界更加模糊,如何将其落实到实证数据中难免成为一大挑战。数字技术在处理复杂数据方面的强大能力,为历史学家应用多征性模型,弥合模型与史料之间的鸿沟提供了强大的技术支持。

这里可以通过社群检测算法在宗教文化史研究中的应用来加以说明。从宋代开始,许多民间信仰从发源地向外传播,在各地信众之间建立起错综复杂的宗教联系,逐渐形成了区域性的宗教文化圈。然而,如何界定和分析这些宗教文化圈却面临诸多挑战。首先,民间信仰的神祇种类繁多,传播路径也非常复杂。既可能存在以水波纹形式从发源地向外逐层扩散的情况,也可能存在随着商人、士大夫的旅行和徙居,在作为商贸、行政和交通枢纽的城市之间跳跃式传播的现象。其次,由于中国社会的多神信仰传统,人神关系呈现出多对多的特点,不仅一个神祇会在多地受到供奉,一个地方也往往存在多种民间信仰并存的情况。史料的不完整性则进一步增加了研究的难度。因此,“宗教文化圈”作为一个分析概念(即模型),在定义上必须保持灵活和开放,多征性定义自然是不二之选。也就是说,一个宗教文化圈必须由一系列民间信仰来定义的。在实证层面,这些信仰未必在圈内的所有地方都出现,也未必仅仅存在于该圈范围内。但是,如何在实证层面落实这一模型就给历史学家带来了巨大的挑战,而社会网络分析中的社群检测算法为解决这个问题提供了一个可行的方案。

尽管社会网络分析已有百年历史,社会科学家也提出过各种与社群检测相关的概念和方法(如 n—cliques、k—clans 等),但这些方法在处理复杂数据方面的效果十分有限。随着计算机技术的发展和算力的提升,学者们开始将迭代算法应用于实证数据,取得了前所未有的成功,伴随新算法出现的 R 语言包和交互式网络可视化软件也大大降低了利用这些算法的门槛。以 2008 年开发的鲁汶算法(Louvain algorithm)为例,该算法对社群的定义十分灵活,只要求每个社群内部的连边密度较高,而社群之间的连边密度较低。该算法还提供了快速有效的迭代方法,首先将网络中的每个节点视为一个独立的社群,然后通过调整邻接节点的社群归属来寻找最优解,亦即通过尽量提升社群内部的连边密度、降低社群之间

① 比如,在《精神疾病诊断与统计手册》(DSM)和《国际疾病分类》(ICD)中,所有精神障碍都是作为多征性概念加以定义的。任何一种特定的精神障碍都是由多个症状定义的,而一个特定个体并不需要具备所有症状就可以被诊断为患有该种精神障碍。只需要观察到一定数量、一定组合的症状就可以被认为符合诊断标准。Robert F. Krueger and Serena Bezdjian, “Enhancing Research and Treatment of Mental Disorders with Dimensional Concepts: Toward DSM-V and ICD-11.” *World Psychiatry*, vol. 8, no. 1, 2009, pp. 3-6。Needham 将多征性概念引入人类学,见 Rodney Needham, “Polythetic Classification: Convergence and Consequences,” *Man* vol. 10, no. 3, 1975, pp. 349-369。有关宗教学研究中应用多征性定义的讨论,见 Jonathan Z. Smith, *Imagining Religion: From Babylon to Jonestown*. University of Chicago Press, 1982, pp. 4—5。

② 《历史学与社会理论》,第 50 页。

的连边密度,实现对给定网络中所有节点的最佳分组。①

在去年发表的一篇论文中,笔者尝试将鲁汶算法应用于宋元时代江南宗教文化圈的研究。笔者利用《宋元方志丛刊》中有关南宋至元代江南地区民间信仰的记录,构建了一个双模网络。该网络包含两种节点,即地点节点(精确到县,共51个)和信仰节点(78个)。两类节点之间的连边表示某信仰是否存在于某地。通过将鲁汶算法应用于这个双模网络,笔者检测出七个社群,每个社群都由数量不等的地点节点和信仰节点构成,可以视为一个区域性的宗教文化圈。这七个社群可以通过矩阵形式直观地呈现,即以矩阵的每一行代表一个地方、每一列代表一个信仰,矩阵的每个元素(取值为“空”或“x”)表示该列信仰是否存在于该行地点。利用鲁汶算法可以将该矩阵的行、列分别切分为七个部分,各自对应鲁汶算法检测出的七个社群(即七个宗教文化圈)。在这个矩阵中,对角线上各区块内的x号密度反映了每个社群内部连边的密度,而对角线以外各区块内的x号密度则反映了社群之间的连边密度,这种数据呈现方式也可以帮助历史学家直观地判断社群检测的合理性。②

在这个例子中可以看出,每个社群都有一系列特征信仰(即被算法归入该社群的那些信仰),这些信仰并不是每个都会出现在该社群内的所有地方;同样,一个地方通常存在多种信仰,而这些信仰也不一定都是该地所属社群的特征信仰。换言之,每个区域性宗教文化圈可以看作一个多征性模型,由一系列信仰(即属性特征)定义。判断一个地方是否可以归入某个宗教文化圈,并不取决于该地是否拥有这个宗教文化圈的所有特征信仰,也不要求该地完全不存在其他宗教文化圈的特征信仰。利用鲁汶算法揭示出来的宗教文化圈,其定义是互递归式的(mutual recursion),亦即一个信仰是否是某个宗教文化圈的特征信仰取决于它是否主要存在于这个文化圈内的各个地点,而一个地点是否属于某个宗教文化圈又反过来取决于此地的信仰是否主要是这个文化圈的特征信仰。对宗教文化圈的这种多征性、互递归式的定义,保持了这一模型的灵活性和开放性,并对史料中可能存在的错误和遗漏提供了较大的包容性。数字技术赋能下的社会网络分析算法,则为历史学家在史料层面落实这种定义提供了可能,为历史学家搭起了连接史料与思维模型的桥梁。值得一提的是,在鲁汶算法加持下,历史学家不需要预先定义一个模型的属性特征。无论是宗教文化圈的数量,还是哪些信仰可以算作哪个宗教文化圈的特征属性,抑或哪些地点属于同一个文化圈,都不是预先设定的,而是通过鲁汶算法在迭代过程中找寻出来的。

数字技术的赋能帮助历史学家突破了传统研究方法的三大限制。③ 首先,如果历史学家仅对某一地方或某一信仰进行个案研究,容易陷入难窥全豹的困境,也就无从判断所研究的个案是否具有代表性。其次,若从个案研究转向整体性研究,即便只涉及51个地点和78个信仰,两者相乘即已存在3978个潜在数据点。要以此为基础来划分宗教文化圈,在宗教文化圈总数不明、特征信仰为何不详、地理边界模糊的多重制约下,单凭直觉和记忆要判断每个信仰和每个地点的归属是极其困难的。如果未来进

---

① Vincent D. Blondel, Jean—Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre, “Fast unfolding of communities in large networks,” *Journal of Statistical Mechanics: Theory and Experiment*, no. 10, 2008: 10008. 鲁汶算法假设网络中的各个社群不重叠,近年发展出来的莱顿算法则通过改进鲁汶算法,将其应用领域进一步扩展到对重叠社群的辨识中。Vincent A. Traag, Ludo Waltman; Nees Jan van Eck, “From Louvain to Leiden: Guaranteeing well—connected communities,” *Scientific Reports*, vol. 9, no. 1, 2019: 5233。

② Song Chen, “Patterns of Integration: A Network Perspective on Popular Religious Connections in China's Lower Yangzi, 1150-1350.” *Religions*, vol. 14, no. 5, 2023, p. 577. 此处所讨论的矩阵,见原文中的矩阵1(Matrix 1)。

③ 这个部分的讨论受到列夫·曼诺维奇(Lev Manovich)等人的启发,他们曾指出数字技术的长处是帮助研究者突破传统研究方式面临的三大制约。尽管此论是就漫画研究而发,但是笔者认为它在史学研究领域同样适用。Lev Manovich, Jeremy Douglass, and Tara Zepel, “How to Compare One Million Manga Pages,” 2011 (https://manovich.net/index.php/projects/how—to—compare).

一步扩展研究的时空涵盖面(如纳入江南以外的数据、探讨从唐宋至明清的变化),更是对人类脑力的极大挑战。再次,即使以传统研究方式能得出类似的研究结果,如何在语言上加以表述也是一大困难。很多信仰在其所属宗教文化圈外存在空间溢出现象,在其所属宗教文化圈内也并非无处不在。不同信仰在空间分布形态上的巨大差异和复杂性,远远超出了自然语言表达的能力。因此,无论是要在史料中落实宗教文化圈这一多征性模型,还是要对相关研究结果进行准确表述,都离不开数字技术的加持。

在此,不妨反思一下数字史学中量化研究的本质。其实,到目前为止,即便在使用量化方法的历史研究中,推论统计的应用也相对较少①,而描述统计仍属主流。推论统计指的是通过抽样、假设检验和区间估计等方法,从有限的样本中推导出有关研究对象总体的结论,而在数字史学研究中,迄今为止,量化分析大多是通过图表和数值指标(如平均数、标准差等)来展示数据的分布情况。例如,前述案例中讨论的矩阵,本质上只是将从宋元方志中提取出来的原始记录(即哪个信仰存在于哪个地方)以矩阵形式呈现,通过调整行、列的位置,将数据分布特征相似的行、列排列在一起。换言之,这种矩阵只是对原始记录的代数表达。它对原始记录进行了简明扼要的概括,并通过调整行列位置,凸显原始记录中蕴含的数据分布特征,从而帮助历史学家从纷繁芜杂的史料向简明抽象的思维模型迈出关键性的一步。这也意味着,这些量化分析结果本身并不构成完整的历史论述,也不具备明确的历史意涵。它们仅仅是在总体层面上对历史记录做出的量化描述,其背后的意义与合理性仍需历史学家结合自身的专业知识和史料中大量无法量化的叙述性内容,通过史料批判、文本细读等传统研究方法进行解读和评估。只有将量化证据与历史诠释相结合,历史学家才能真正实现从史料到模型的跨越,提出具有说服力的历史论述。

语境转换与数据关联

如果数字史学中的量化分析结果仅仅是对历史记录中部分内容的数值化描述,那么它描述了哪些内容,又舍弃了哪些内容呢? 这里我们仍从上述案例出发来看一个例子。

> 《嘉定赤城志》卷三一(祠庙门)
>
> **临海**
>
> ……
>
> 灵康庙:在县东南二十里白鹤山,祀东汉**赵炳**。炳,东阳人,能为越方,善禁咒。与闽人徐登遇于乌伤溪上,各试其术,以起人疾。登死,炳入章安,神幻事甚众,后为章安令华表所害。世传其尸泝流止今处,故祠之,以登侑食。入国朝,迭显灵异。元丰七年,赐今额。崇宁三年,封仁济侯。大观二年,进显仁公。政和三年,进灵顺王。宣和四年,加显佑。建炎四年,加广惠。庆元二年,加善应。开禧三年,改善应为威烈。宣和二年,庙毁于寇,寇退复建。淳熙八年唐守仲友、嘉定六年俞守建,皆更新之。【夹行】唐守记云:[略]。章守得象题云:[略]。②

① 目前,使用推论统计的历史研究成果大多数是社会科学家完成的,如 Fangqi Wen, Erik H. Wang, and Michael Hout, "Social Mobility in the Tang Dynasty as the Imperial Examination Rose and Aristocratic Family Pedigree Declined, 618—907 CE." *Proceedings of the National Academy of Sciences — PNAS*, vol. 121, no. 4, 2024, pp. e2305564121-e2305564121. Joy Chen, Erik H. Wang, Xiaoming Zhang, "From Powerholders to Stakeholders: State—Building with Elite Compensation in Early Medieval China," *American Journal of Political Science*, 2024, https://doi.org/10.1111/ajps.12888. Joy Chen and Erik H. Wang, "State—Building or State—Weakening? The Consequences of Military Control in Medieval China," June 29, 2024, available at SSRN: http://dx.doi.org/10.2139/ssrn.4880900. Yuhua Wang, "Blood Is Thicker Than Water: Elite Kinship Networks and State Building in Imperial China," *The American Political Science Review*, vol. 116, no. 3, 2022, pp. 896-910. Ying Bai, Ruixue Jia, Jiaojiao Yang, "Web of Power: How Elite Networks Shaped War and Politics in China," *The Quarterly Journal of Economics*, vol. 138, no. 2, 2023, pp. 1067-1108.

② 《宋元方志丛刊》第 7 册,北京:中华书局 1990 年版,第 5719 页(《嘉定赤城志》卷 31,第 8 页上下)。

这段史料可谓翔实,介绍了灵康庙的地理方位、建庙缘由,详述了所祀神祇的身分与神迹,历数了两宋朝廷的褒崇之政,还收录了淳熙八年知州唐仲友的碑记全文(即“唐守记”)。这种碑记堪称弥足珍贵的史料,在华南学派“进村找庙、进庙找碑”的历史人类学研究传统中尤其受到重视。但前例所述的双模网络分析中,仅提取了“临海县—赵炳信仰”一条数据,其余内容概行舍弃。这种研究方式的局限自不待言。例如,它忽略了在人类学传统下研究民间信仰的学者通常会特别留意的一些议题,比如供奉同一神祇(如“赵炳”)的信众是否会对这个宗教符号的意涵做出完全不同的诠释、是否会采用大相径庭的科仪等。从这个角度来看,宋元方志中的历史记录为理解这些民间信仰提供了丰富的文化语境,而当我们将“临海县—赵炳信仰”这条信息作为关系数据抽取出来时,原有语境也就被排除于后续的量化研究之外了。

然而,如果因此而否定数字史学或量化研究的意义,恐怕也有失公允。因为在舍弃原有语境的同时,数字史学也会从不同的研究议题和研究方法出发,构筑起新的语境,并在新语境下重新审视每条数据的历史意义。比如,在前例的双模网络研究中,笔者研究的并非“临海县—赵炳信仰”这条关系本身,而是把它与“临海县—张王信仰”“临海县—陈果仁信仰”“庆元府象山县—赵炳信仰”等其他数百条从《宋元方志丛刊》中提取出来的关系数据结合起来,把这些关系的集合作为一个整体,进而研究其背后蕴含的结构性特征。① 正是在这个整合了各地信仰数据的新语境中,历史学家才能利用算法在实证层面落实宗教文化圈的概念,才能分析信仰分布的宏观空间形态,揭示各个信仰、各个地方的相互关联,并在这种宏观背景下从全新的视角理解不同信仰(如局限于台州一带的赵炳信仰、局限于徽州一带的汪华信仰、或者分布范围更广的张王、陈果仁等信仰)在界定、联结不同宗教文化圈方面所扮演的不同角色。

数字史学为历史学家提供了一种与史料互动的新方式。在数字史学的研究范式下,历史文献不仅是可以被细读、被互校的对象,也是一种蕴含着丰富链接的数据宝藏,文献之间有着内在的呼应关系。数字技术用结构化、标准化的方式把这些数据从叙述性文字中提取出来,在脱离但不背离原有语境的情况下进行关联和呈现,进而揭示在文献细读时不易察觉的历史现象,把抽象的思维模型落实在历史证据中。这种数据提取和关联,打破了原有语境,但也是基于史料之间内在呼应关系的思维重构,可以帮助历史学家在一定程度上绕开史料书写者的视角,穿透纷繁的文字表象,更深入地触摸到史料中蕴藏的历史实态。

对从多种史料中提取出来的数据进行关联,正是关系型数据库的核心价值,也是数字基础设施建设的重要内容。这里以传记数据库为例就数据库建设和数据整合对历史研究的价值加以探讨。数字史学的研究周期涵盖了三个阶段:从选定的史料中抽取与研究课题相关的数据、对不同史料中抽取的同类数据进行消歧和整合、以及对整合后的数据进行整体性研究。这一周期对应了数据挖掘、数据库建设、数据分析三个研究阶段,也是一个将数据抽离原有语境、通过数据关联建立新语境、最后在新语境下重新审视数据的过程。这些特点与群体传记学研究(prosopography)若合符节。根据劳伦斯·斯通(Lawrence Stone)的定义,群体传记学是通过对一群人物生平进行整体性研究,来探讨他们共同的背景特征。在研究过程中,要先确定研究范围,然后提出一系列整齐划一的问题(如有关人物居住地、家庭背景、社会出身、个人财富的数量与来源、教育程度、职业生涯、宗教信仰等方面),最后通过排比和整合每个人物的相关信息来找寻具有显著性的变量。② 进入 21 世纪后,社会网络分析方法引入群体传记学领域,开启了新的研究方向,即将研究者的关注点从探索一个历史人群的共同背景特征,转向研究其内

---

① 参见陈松、赵薇:《从隐喻到模型——作为研究和批评路径的网络分析》,《数字人文》2022 年第 1 期,第 1—13 页。

② Lawrence Stone, “Prosopography,” *Daedalus*, vol. 100, no. 1, 1971, p. 46.

部相互关系和交往模式的整体结构。① 不论是哪种类型的群体传记学研究,由于人物的生平信息通常散见于不同的史料中,研究者都需要将这些分散的信息从原有语境中提取出来,然后汇整为一个可供量化分析的数据集。对于侧重发掘人际关系中结构性特征的新式群体传记学,数据整合尤为重要。

例如,笔者在早年研究中曾使用中国历代人物传记资料库(CBDB)的数据,重建了在 1040 年代和 1210 年代担任郡守的宋代官员之间的亲属关系,指出 1040 年代的郡守至少五分之二通过血缘或姻缘联成一张庞大的关系网,而这一现象到 1210 年代则不复存在。② 以 1040 年代的亲属网络为例,该网络由 211 个人物(节点)和 757 条亲属关系(连边)组成。其中众多亲属关系在史料中并无直接表述,必须通过散见于多篇墓志中的父子关系和婚姻关系拼凑出来。比如,同于 1040 年代担任知州的欧阳修(1007—1072)和庞籍(988—1063)之间的姻亲关系,必须通过散见于下列四方墓志的父子关系和夫妻关系进行推导:

> 欧阳伯和墓志铭(欧阳发):"君欧阳氏,讳发,字伯和,庐陵人,太子少师文忠公讳修之长子也。……夫人吴氏,故丞相正宪公充之女,封寿安县君。"
>
> 吴正宪公充墓志铭(吴充):"皇考待问,登咸平进士第,官至尚书礼部侍郎……女四:长适殿中丞欧阳发……"
>
> 资政殿大学士尚书左丞赠吏部尚书正肃吴公墓志铭(吴育):"父讳待问,官至礼部侍郎,赠太保;……女三人:长适集贤校理韩宗彦,次适著作佐郎庞元英,皆早卒……"
>
> 太子太保庞公墓志铭(庞籍):"男五人:长曰元鲁……次元英,太常博士……"③

从上述墓志中,可以提取出欧阳发—欧阳修、欧阳发—吴充、吴充—吴待问、吴育—吴待问、吴育—庞元英、庞籍—庞元英六条亲属关系,将这些关系拼接起来便能得出欧阳修之子和庞籍之子同为吴待问孙婿的结论。④ 虽然历史学家通过熟读史料也许可以做到这一点,但数字技术大大提高了数据挖掘和人名消歧的效率,以此为基础建立的关系型数据库(即 CBDB),更为数据分享和研究复现提供了可能。

正因数据关联对群体传记学有着重大价值,近年来,这一领域的累累硕果无不是以大型数据集的建设为前提的。仅就中国史领域而言,谭凯(Nicholas Tackett)从数千方墓志铭中提取亲属关系数据,重建了九世纪首都精英的婚姻关系网。虽然他没有使用社群检测算法进行量化分析,但仅凭可视化已在实证层面上有效地落实了山东士族和关陇士族这两个唐史学者常常讨论的概念。⑤ 马德伟(Marcus Bingenheimer)对早期佛教网络的研究也同样以中国佛教历史社会网络数据集(HSNCB)为基础,该数据集从历代高僧传等佛教传记资料中抽取并整合了一万七千多人之间的多种社会关系,涵盖了师徒传承、合作译经等诸多面向。⑥ 李中清团队建设的中国历史官员量化数据库(清代)、中国大学生数据库更为

---

① Dion Smythe, "Prosopography," in *The Oxford Handbook of Byzantine Studies*, ed. Robin Cormack, John F. Haldon, and Elizabeth Jeffreys, New York: Oxford University Press, 2008, pp. 176—181。包弼德、王宏苏、傅君劢、陈松、柳舟、朱厚权:《"中国历代人物传记资料库"(CBDB)的历史、方法与未来》,《数字人文研究》2021 年第 1 期,第 23 页。

② Song Chen, "Governing a Multicentered Empire: Prefects and Their Networks in the 1040s and 1210s," in *State Power in China*, 900—1325, edited by Patricia Buckley Ebrey and Paul Jakov Smith (Seattle: University of Washington Press, 2016), pp. 101-152.

③ 曾枣庄、刘琳编:《全宋文》第 128 册,卷 2769,第 126 页;第 79 册,卷 1718,第 54 页;第 35 册,卷 753,第 349 页;第 56 册,卷 1226,第 277 页。

④ 参见包弼德、王宏苏、傅君劢、陈松、柳舟、朱厚权:《"中国历代人物传记资料库"(CBDB)的历史、方法与未来》,《数字人文研究》2021 年第 1 期,第 23 页图 1。

⑤ 谭凯:《中古中国门阀大族的消亡》,北京:社会科学文献出版社 2017 年版,第 3 章。

⑥ 马德伟:《历史社会网络分析在中国佛教——道安、慧远、鸠摩罗什的案例》,《佛学研究》2023 年第 1 期,第 179—208 页。

研究清代官僚、民国以后的大学生提供了坚若磐石的数据支持。在近现代史方面,李雷雷(Thorben Pelzer)对 20 世纪上半叶中国技术官僚的研究亦得益于其团队历时多年建成的中国工程师关系型数据库(CERD)。① 这些数据库通过汇整和关联从墓志铭、僧尼传、缙绅录、学籍档案、专业团体会员录等多种史料中提取的传记信息,为研究者在群体层面对历史人物开展量化研究提供了丰富的基础数据,为研究者从历史人群的共同特征、相互关系、互动模式出发重新审视单个人物的人生轨迹及其历史意涵提供了可能。

显然,数据消歧、数据关联和数据库建设对历史研究的价值绝不仅限于群体传记学研究。如前文所述的宋元江南宗教文化圈研究也会遇到数据消歧、数据关联等问题。比如,同为赵炳信仰,在台州诸县则以灵康庙或灵康行祠的形式出现,而他在庆元府象山县的神祠则被称为白鹤庙。反过来,同称英烈庙,在常州宜兴县奉祀的是周处,在严州淳安县奉祀的则为钱嵒钱嵩兄弟。因此,要在更大的时空维度上开展宗教文化圈的量化研究或其他类似研究,现有的方志全文数据库就难以满足研究需求了,势必要依赖以地方志为基础的各种专题数据库的建设。在这方面,魏希德团队开展的中国物质基础设施计划代表了一个极有价值的学术方向。②

除了建设大型数据库之外,世界各地的学术团队也相继开发出玛库斯(MARKUS)以及与之配套的 COMARKUS、IMMARKUS 等标注平台③、DocuSky④、“吾与点”⑤、HistText⑥、LoGaRT⑦ 等通用型或专题型研究平台,或像中国基督教历史资料库⑧那样从项目伊始就将数据库和可视化工具融为一体,大大降低了人文学者从史料中识别和抽取命名实体及其相互关系、开展可视化和数据分析的技术门槛。

语境转换与文本分析

建设关系型数据库不是实现语境转换、桥接史料与模型的唯一方式。有些数字史学研究项目使用无监督学习的方法,在不依赖标注数据或结构化数据库的情况下,将历史文本直接拿来作为量化分析的输入项,使用文本复用、词频分析、主题模型、文档聚类等算法揭示文本的主题和风格。

文本复用是指在不同文本中识别和提取相似或重复的内容,为分析文本之间的关联性提供数据支持。这里以德龙(Donald Sturgeon)对先秦两汉文献的聚类分析为例,探讨文本复用如何帮助历史学家实现语境转换和模型落地。⑨ 德龙的研究分为两步。在第一步中,德龙使用字符串匹配等算法识别不同文本中的相似部分(即复用文本),如下例中的黑体字部分:

> 《荀子·劝学》:吾尝终日而思矣,不如须臾之所学也。吾尝跂而望矣,不如登高之博见也。登高而招,臂非加长也,而见者远;**顺风而呼,声非加疾也,**而闻者彰。假舆马者,非利足也,而致千里;假舟楫者,非能水也,而绝江河。君子生非异也,善假于物也。

---

① Thorben Pelzer, *Engineering Trouble?: US—Chinese Experiences of Professional Discontent*, 1905—1945. Leiden; Boston: Brill, 2023. 李氏所用数据库,见 Thorben, Pelze et al., "Chinese Engineers Relational Database (CERD)," Geneva, Budapest: Zenodo, 2022 (http://doi.org/10.5281/zenodo.4075601)。

② https://www.infrastructurelives.eu/

③ https://dh.chinese—empires.eu/markus/beta/

④ https://docusky.org.tw/DocuSky/home/

⑤ https://kvlab.org/project/wyd/

⑥ https://bookdown.enpchina.eu/rpackage/HistTextRManual.html

⑦ https://www.mpiwg—berlin.mpg.de/research/projects/logart—local—gazetteers—research—tools

⑧ https://chcdatabase.com/zh/

⑨ Donald Sturgeon, "Unsupervised Identification of Text Reuse in Early Chinese Literature," *Digital Scholarship in the Humanities*, vol. 33, no. 3, 2008, pp. 670-684.

《吕氏春秋·慎大览·顺说》:善说者若巧士,因人之力以自为力;因其来而与来,因其往而与往;不设形象,与生与长;而言之与响;与盛与衰,以之所归;力虽多,材虽劲,以制其命。**顺风而呼,声不加疾也**;际高而望,目不加明也;所因便也。……役人得其所欲,己亦得其所欲。以此术也,是用万乘之国,其霸犹少,桓公则难与往也。

第二步,德龙以每对文本中复用部分的比例为依据,构建了文本之间的加权网络。这个加权网络以文本之间的相似性为基础,文本之间的连边权重即由复用比例决定。结合可视化和社群检测算法,德龙对这个网络中的文本进行了分类。

虽然《荀子·劝学》和《吕氏春秋·慎大览·顺说》都使用了“顺风而呼,声[非/不]加疾也”这段文字,但是结合上下文的语境可以看出两者在意涵上有很大差别。《荀子》以此为喻,是为了反对思孟学派对“思”的偏重,而提出“学”的重要性,主张“借助于自我以外的事物来扩展、延伸或者提升自己”(即所谓“善假于物”),而不是闭门而思。① 而《吕氏春秋》则是为了说明说服别人的技巧在于“因人之力以自为力”,也就是要避免针锋相对,要善于使用引导、劝诱的方式来申论自己的主张,达到说服对方的目的。德龙在量化研究的第一步就将复用文本(“顺风而呼,声[非/不]加疾也”)从原文语境中抽离出来了。这种做法固然有其缺陷,但是文本复用算法对“形式相似”(formal similarity)的关注也促使我们注意到《荀子》和《吕氏春秋》之间共享的知识资源。不仅如此,在第二步中,德龙更将这段复用文本(“顺风而呼,声[非/不]加疾也”)与算法识别出来的其他复用文本相结合,以此为基础构建出新的语境(即文本相似度网络)、对文献进行分类,使研究者可以在转换后的新语境下评估每部典籍在整个先秦两汉文献传统中的地位。

值得一提的是,与前述的宗教文化圈研究一样,德龙得出的文献类别不是预设的,而是算法根据文本复用的程度自动发现的。德龙基于自己的专业素养为每类文献贴上如字书、医书、法家、道家等标签,从而将每个标签所对应的文献类别从抽象的思维模型落实到具体的实证证据上。这个过程并不排斥研究者的主体性,相反,最终的研究成果正是研究者、算法和实证数据之间频繁互动的产物。研究者不仅在为文献类别贴标签时加入了自己对分类结果的诠释,而且在整个研究过程中,研究者的专业判断无时不在。在诸如如何设定文本复用的临界值、使用文本复用程度来衡量文本相似度是否合理等问题上,研究者都需要根据自己的知识和经验做出选择。

在文本复用之外,还有很多文本分析是以词频和词语共现为基础的。词频分析通常采用“词口袋”法,即完全忽略句法规则,将每篇文本视为简单无序的词语集合,先将文本拆解为一个个词语,然后计算每个词语出现的次数(即词频),最后根据高频词来推断文本的内容和风格。以词频分析为基础,学者们还进一步分析词语的共现情况,发展出了主题建模等更为精细的算法。由于同一个词语在不同语境下可能具有不同的含义(反之,不同词语也可能表达相似的含义),因此主题建模使用概率模型,通过分析词频和词语共现来推断语义,自动识别文本中的潜在主题。

词频分析是如何创造新语境和促成模型落地的呢?这里以笔者对宋代州县学碑记的分析略加说明。② 这项研究与德龙对先秦两汉文献的分析类似,同样分为识别特征词汇和进行文本聚类两个步骤。具体来说,在第一步中,笔者从双字组合(2—gram)出发在每篇文本中寻找有意义的古汉语词汇,然后综合考虑每个词汇在给定碑记中的出现频率(TF)及其在整个语料库中的稀有程度(IDF),据此为每篇碑记中的每个词汇分别赋予权重(即 TF—IDF 值)。这个权重综合反映了该词汇在每篇碑记及整个语料库中的重要性。经过这番处理后,每篇碑记被转换为一个多维向量,向量的每个维度对应一个词汇的权重。每篇碑记在这个特征空间中的位置反映了它在主题和风格方面与其他碑记的相似程度。在第二

① 王博:《论〈劝学篇〉在〈荀子〉及儒家中的意义》,《哲学研究》2008 年第 5 期,第 63—64 页。

② 陈松:《为学作记——从网络分析和文本分析视角看宋代地方官学碑记的作者与主题》,《数字人文》2021 年第 1 期,第 24—72 页。

步中,笔者对这些向量(即碑记)进行降维处理,并使用K—均值算法加以聚类,亦即根据向量之间的距离确定哪些碑记可以归入同一类别。这一研究流程实际上也是拆解原有语境、再为每篇碑记重建新语境的过程。试看朱熹在淳熙四年(1177)撰写的《静江府学记》:

> 古者圣王设为学校,以教其民,由家及国,大小有序,使其民无不入乎其中而受学焉。而其所以教之之具,则皆因其天赋之秉彝而为之品节,以开导而劝勉之,使其明诸心、修诸身,行于父子、兄弟、夫妇、朋友之间,而推之以达乎君臣上下、人民事物之际,必无不尽其分焉者。及其学之既成,则又兴其贤且能者寘之列位。是以当是之时,理义休明,**风俗**醇厚,而公卿、大夫、列士之选无不得其人焉。此**先王**学校之官所以为政事之本、道德之归而不可以一日废焉者也。
>
> 至于**后世**学校之设,虽或不异乎**先王**之时,然其师之所以教,弟子之所以学,则皆忘本逐末,怀利去义而无复**先王**之意。以故学校之名虽在,而其实不举,其效至于**风俗**日敝,人材日衰,虽以汉唐之盛隆,而无以仿佛乎三代之叔季。然犹莫有察其所以然者,顾遂以学校为虚文而无所与于道德政理之实,于是为士者求道于老子、释氏之门,为吏者责治乎簿书期会之最,盖学校之仅存而不至于遂废者,亦无几耳。
>
> 乃者圣上慨然悯其如此,亲屈銮路,临幸学宫,发诏诸生,励之以为君子之儒而无慕乎人爵者,德意既甚美矣。而静江守臣广汉张侯栻适以斯时一新其府之学,亦既毕事,则命其属具图与书,使人于武夷山间谒熹文以记之。(后略)①

这段文字的撰写缘于朱子友人张栻的请托。时知静江府(今广西桂林)的张栻重修了本府官学,毕事之际专门派人前往武夷山请朱子作记。节录的段落偏重议论,在强调细读的传统史学研究中,历史学家通常会从上下文的语境出发,并结合朱张两人的思想背景进行解读。朱子的立论框架是“古者”与“后世”学校的区别。在朱子看来,先王设立学校是为了“因其天赋之秉彝而为之品节,以开导而劝勉之”,从而实现每个人的人格养成,“使其明诸心、修诸身,行于父子、兄弟、夫妇、朋友之间,而推之以达乎君臣上下、人民事物之际,必无不尽其分焉者”;而后世学校“则皆忘本逐末,怀利去义”,其名虽在,其实不举。简言之,这段文字重新定义了官学的根本宗旨,即要根据每个生徒的秉彝加以开导劝勉,使其心明、使其身修,而“道德政理”无不是从此生发出来的。② 天赋、明心、修身、本末、义利等概念处处体现出朱子理学的思想脉络。

使用TF—IDF值寻找特征词汇则为理解这段文字提供了完全不同的视角,揭示了这篇碑记与11世纪后半叶思想潮流的承继性。通过分析这篇碑记中常见的双字组合,计算机算法将研究者的注意力引向“学校”“君子”“先王”“后世”“风俗”“道德”等特征词汇,将这些词汇从原文的语境中抽离出来,置于一个由全部773篇现存宋代州县学碑记中提取出来的特征词汇构成的新语境中进行解读。在这个新语境中,每个词汇的重要性(TF—IDF)不仅取决于它在单篇碑记中的角色,而且综合考虑了它在整个语料库中的独特性。基于K—均值算法的文本聚类便是以这种解读为基础进行分析的结果,它将现存宋代州县学碑记分为三类,而朱熹的《静江府学记》被归入第二类。这一类学记主要流行于11世纪中期到12世纪末,其典型特征是大量使用“先王”“后世”“风俗”“道德”等词汇,而这些词汇在学记中频繁出现的时间大致集中在11世纪下半叶(表1)。换言之,虽然朱子在这篇学记中强调官学教育应以明心修身为鹄的,但他也接纳了先王兴学是要为朝廷育才(即实现“公卿、大夫、列士无不得其人”)这个大前提。他在讲明心修身之时,将自己的主张嵌入先王后世之变的论述框架中,而这个框架实际上承自11世纪中期的思想传统。换言之,算法提取的特征词汇为理解朱子的《静江府学记》提供了一种更加宏阔的知识背景,这也是在观察个篇时不易把握、但在当时人看来或许不言自明的思想传统。由于算法更

① 曾枣庄、刘琳编:《全宋文》,第252册,第5653卷,第73页。

② 这与朱子在他处表达的思想是完全一致的。见黎靖德编:《朱子语类》,北京:中华书局1986年版,卷4,第58页;卷13,第231页。

容易识别出不同篇章之间的内在呼应关系,而细读时研究者更倾向抓住朱子的思想倾向、对他提出的新见沉潜发微,历史学家往往要两相参照,才能更加全面地看出朱子与时代的互动来。

表 1 第二类学记特征词汇的历时分布

学记年代 \ 特征词汇	含有特征词汇的州县学碑记篇数或比例					
	先王（N）	先王（%）	后世（N）	后世（%）	风俗（N）	风俗（%）
980—999	1	16.7	0	0.0	1	16.67
1000—1019	2	12.5	2	12.5	5	31.25
1020—1039	1	6.3	2	12.5	3	18.75
1040—1059	5	10.2	12	24.5	11	22.45
1060—1079	14	41.2	8	23.5	7	20.59
1080—1099	14	27.5	15	29.4	10	19.61
1100—1119	7	25.0	4	14.3	9	32.14
1120—1139	4	14.8	6	22.2	3	11.11
1140—1159	10	11.9	17	20.2	11	13.10
1160—1179	11	13.3	18	21.7	14	16.87
1180—1199	6	6.5	13	14.0	11	11.83
1200—1219	9	9.5	16	16.8	10	10.53
1220—1239	7	7.5	14	15.1	6	6.45
1240—1259	0	11.8	4	6.6	2	3.28
1260—1279	0	0.0	1	2.8	4	11.11
总计	91	11.8	132	17.1	107	13.86

在数字基础设施方面,以文本分析为核心的研究项目所需的技术支持有别于关系型数据库或玛库斯等文本标注工具。文本研究能否顺利开展通常取决于两个要素:一是可以用作输入数据的全文资料库,二是便于人文学者使用的文本分析工具。全文数据库的建设已有数十年的历史,学术界在古籍数字化、档案材料数字化以及 OCR 技术等方面取得了显著进展。然而,目前许多全文数据库只支持关键词检索,而文本复用、词频分析、主题建模等算法都需要访问大量文献的完整文本。近年来,如何在不侵犯出版商版权的前提下实现这一功能是众多学者面临的挑战。在这一方面,德国马克斯·普朗克科学史研究所付出的努力具有开创性的意义。他们开发的 RISE 软件(“欧亚研究基础设施”)通过名为 SHINE 的一组标准化应用编程接口(API),实现了第三方研究工具与多种第三方文献资源之间的安全链接,使研究者在没有全文下载权限的情况下仍能使用数字分析工具访问受版权保护的文献资源。① 最近,该所更将大量史籍制作成 N—gram 数据集(即单字、双字组合、三字组合)供下载使用,目前已涵盖《续修四库全书》、《道藏辑要》和大约 8000 种地方志等海量文献。这种做法不仅平衡了版权限制与数字人文研究的需求,还有效避免了文本分析过程中拆分 N—gram 的重复劳动。在工具建设方面,中国哲学书电子化计划先前已经开发了一系列插件,提供了生成 N—gram、计算 TF—IDF、测量文本相似度和执行 PCA 分析等功能。② 最近 TextPAIR Viewer (TPV)平台的开发,进一步降低了人文学者识别复用文本和以可视化方式考察文本关系的技术门槛。③

近年,人工神经网络的发展也给自然语言处理领域带来了翻天覆地的变化,为文本分析开拓出新的前景。相较于通常以专名形式(如人名、地名、职官等)出现的命名实体,社会关系、情感体验、心理特质

① “RISE and SHINE: Research Infrastructure for the Study of Eurasia” (https://www.mpiwg—berlin.mpg.de/research/projects/rise—and—shine—research—infrastructure—study—eurasia)

② https://ctext.org/tools/plugins/list/zh

③ https://users.rcc.uchicago.edu/~jcarlsen/TPV/TPV_histories/

等表述方式千变万化,对使用数字技术识别和抽取带来了巨大挑战,因此,长期以来很多研究不得不止步于使用共现数据。① 以人工神经网络为架构开发出来的 Transformer 模型在解决这些问题方面已展现出强大的潜力。自 2017 年推出以来,促成了 BERT、Sentence—BERT(SBERT)、GPT 等预训练模型的发展。在自然语言处理领域中,这些模型的应用日益广泛,在很多任务上的表现都优于早年的词嵌入模型(如 word2vec)和长短期记忆(LSTM)等循环神经网络(RNN)模型。这些预训练模型虽然是使用维基百科等大型语料库进行训练的,但可以针对特定任务进行微调。和先前的词嵌入模型不同,这些预训练模型可以对一个整句或一整段文字进行嵌入,因此它们生成的句嵌入(sentence embeddings)不是仅仅基于单个词语的含义,而能考虑到整句话或整段文字语境,从而更好地捕捉文本之间的语义相似度(semantic similarity)。上述特性使这些预训练模型在中国史研究领域得到日益广泛的运用,使文本分析在原有的 TF—IDF、主题建模等基础上更进一步。如马德伟等人在研究公元 500 年至 800 年间的汉文佛典时,先使用以殆知阁文献为基础训练出来的 GuwenBERT 生成句嵌入,再使用 BERTopic 进行主题建模,揭示了本土佛教文献与汉译佛典在主题上的差别。② 在陈钰琪及其合作者的研究中,她们使用预训练模型将心理构念(psychological constructs)和价值观念从史料中抽取出来,置于更宏观的历史语境下予以解读。在一项具有开创意义的研究中,陈钰琪等将语境化构念模型(CCR)应用于中国历史文献,以句嵌入为基础计算不同文本之间语义相似度。她们使用引文推荐模型(QuoteR)根据现代心理学量表从中文古籍中寻找相似度高的引文,经过人工筛选,最终得到由古籍引文构成的文言文量表。随后,她们使用 SBERT 对文言文量表和自建的中国历史心理语料库(C—HI—PSY)分别进行句嵌入表征,然后通过计算两组向量表征之间的余弦相似度,完成对语料库的心理学分析。她们的研究表明,和主流心理学界采用的基于预定义词表和词嵌入模型的分布式词典表征方法(DDR)相比,基于句嵌入模型的算法取得了更好的效果。于是,她们使用中国历史心理语料库对 SBERT 模型进行微调,并将上述研究方法应用于宋代史料,从宏观的视角观察北宋熙丰变法时期不同官员在价值观念上的差异,以此解释他们对新法的不同立场。③ 在另一项刚刚起步的研究中,陈钰琪等还探讨使用 GPT 构建生成式人工智能体(generative AI agents)来模拟魏晋清谈中的不同角色,并通过使用不同时代的史料作为训练语料,揭示中国历史上思想观念的变迁。

结　语

历史研究离不开史料,也离不开作为历史学家思维建构的模型。模型对过去发生的事件加以简化和抽象,给过去赋予意义。如何跨越横亘于史料与模型之间的鸿沟,乃是所有历史研究都无法回避的问题。尤其是多征性模型,在提供了更加灵活的分析框架的同时,也带来了边界模糊、数据繁杂、识别困难等问题。数字技术以其强大的数据处理能力,为历史学家将模型落实于史料之中提供了前所未有的可能。

数字技术的应用也深刻改变了历史学家与史料互动的方式。它引导历史学家从单篇文献的固有语境中跳脱出来,关注史料之间内在的呼应关系,将史料视为蕴含着丰富链接的数据宝藏,通过数据提取

---

① 使用共现数据进行研究为后续的历史诠释带来了巨大的挑战,不过也有一些较为成功的案例,如 Hilde De Weerdt, Brent Ho, Allon Wagner, Jiyan Qiao, and Mingkin Chu, “Is There a Faction in This List?” *Journal of Chinese History*, vol. 4, no. 2, 2020, pp. 347-389。

② Marcus Bingenheimer, Justin Brody, and Ryan Nichols, “Experiments in Distant Reading: Using Topic Modeling on Chinese Buddhist Texts from 500—800 CE,” *Digital Humanities Quarterly*, forthcoming.

③ Yuqi Chen, Sixuan Li, Ying Li, and Mohammad Atari, “Surveying the Dead Minds: Historical—Psychological Text Analysis with Contextualized Construct Representation (CCR) for Classical Chinese,” https://arxiv.org/abs/2403.00509.

和数据关联为历史研究搭建起更为宏观、更为结构化和标准化的新语境,进而揭示在文献细读时不易察觉的历史现象。这种语境转换,本质上是基于史料之间内在呼应关系的思维重构,它为历史学家穿透纷繁的文字表象,更深入地发掘那些普遍性、典型性的历史现象提供了新的契机。它引导历史学家将个案研究与整体性研究相结合,启发历史学家从更加宏观的、统合的视角重新审视个别现象的历史意义,拓宽了史学研究的视野,也拓展了史学研究的范式。但是,数字史学并不是传统史学的替代,在未来的研究中要如何有效地整合宏观和微观视角、结合定性与定量研究仍有待历史学家在研究实践中继续探索。

Grounding the Models and Recontextualizing the Sources: Historical Research Empowered by Digital Technology

Chen Song

(*Department of East Asian Studies, Bucknell University, USA*)

Abstract: No historical research is possible without sources and models. Models are historians' intellectual constructs that give meaning to past events through simplification and abstraction. How to bridge the gap between sources and models is a challenge that all historians must face. Digital technology, with its powerful data-processing capabilities, offers unprecedented opportunities for grounding models in historical sources. The adoption of digital technology in historical research has profoundly transformed how historians engage with sources. It guides historians to think beyond the original context of individual source materials and facilitates the construction of a more macroscopic and more structured context within which patterns that elude close reading become discernible. This new context, constructed by extracting and linking data from diverse historical sources, is a manifestation of the inherent connections among these sources. As a result, digital technology broadens the scope and paradigms of historical research, encouraging historians to combine case studies with more holistic approaches and to reassess the significance of individual historical phenomenon from more global and integrated perspectives.

Key words: model-grounding; recontextualization; quantitative history; digital history; digital humanities

(责任编辑　郑　园)